Applications in Plant Sciences

APPLICATION ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Green Digitization: Online Botanical Collections Data Answering Real-World Questions

Species distribution modeling based on the automated identification of citizen observations

Christophe Botella[1,2,3,4], Alexis Joly[1], Pierre Bonnet[3,5,7], Pascal Monestiez[4], and François Munoz[6]

Manuscript received 8 September 2017; revision accepted 2 January 2018.

[1] Institut national de recherche en informatique et en automatique (INRIA) Sophia-Antipolis, ZENITH team, Laboratory of Informatics, Robotics and Microelectronics–Joint Research Unit 5506-CC 477, 161 rue Ada, 34095 Montpellier CEDEX 5, France

[2] Institut National de la Recherche Agronomique (INRA), Joint Research Unit Botanique et modélisation de l'architecture des plantes et des végétations (UMR AMAP), F-34398 Montpellier, France

[3] AMAP, Université de Montpellier, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), French National Center for Scientific Research, INRA, IRD, Montpellier, France

[4] BioSP, INRA, Site Agroparc, 84914 Avignon, France

[5] CIRAD, UMR AMAP, F-34398 Montpellier, France

[6] Université Grenoble Alpes, Laboratoire d'Écologie Alpine, CS 40700, 38058 Grenoble CEDEX, France

[7] Author for correspondence: pierre.bonnet@cirad.fr

Citation: Botella, C., A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. 2018. Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences* 6(2): e1029.

doi:10.1002/aps3.1029

PREMISE OF THE STUDY: A species distribution model computed with automatically identified plant observations was developed and evaluated to contribute to future ecological studies.

METHODS: We used deep learning techniques to automatically identify opportunistic plant observations made by citizens through a popular mobile application. We compared species distribution modeling of invasive alien plants based on these data to inventories made by experts.

RESULTS: The trained models have a reasonable predictive effectiveness for some species, but they are biased by the massive presence of cultivated specimens.

DISCUSSION: The method proposed here allows for fine-grained and regular monitoring of some species of interest based on opportunistic observations. More in-depth investigation of the typology of the observations and the sampling bias should help improve the approach in the future.

KEY WORDS automated species identification; citizen science; crowdsourcing; deep learning; invasive alien species; species distribution modeling.

Identifying organisms is a key step in accessing information related to the ecology of species. Specifically, large-scale monitoring of species distribution dynamics is essential in the context of global change. Such monitoring requires intensive occurrence data, but such data are lacking due to the level of expertise necessary to correctly identify and record living organisms. This is especially true for plants, which are one of the most difficult groups to identify, with more than 350,000 known species on earth. The Rio Conference of 1992 (the Earth Summit, United Nations Conference on Environment and Development [UNCED], Rio de Janeiro, Brazil, 3–14 June 1992 [http://www.un.org/geninfo/bp/enviro.html]) recognized this taxonomic gap as a major obstacle to the global implementation of the Convention on Biological Diversity. Gaston and O'Neill (2004) discussed the potential of using automated identification approaches, typically based on machine learning and multimedia data analysis methods, to produce more intensive occurrence data. They suggested that if the scientific community is able to (1) overcome the production of large training data sets, (2) more precisely identify and evaluate error rates, (3) scale up automated approaches, and (4) detect novel species, it will then be possible to initiate the

development of a generic automated species identification system. Such a system should then open important opportunities for studies in biology, ecology, and related fields.

Since Gaston and O'Neill (2004) raised the question, enormous work has been done on the development of automated approaches for plant species identification (Casanova et al., 2009; Yanikoglu et al., 2014; Lee et al., 2015; Champ et al., 2016; Goëau et al., 2016; Joly et al., 2016; Wilf et al., 2016; Wäldchen and Mäder, 2017). Deep learning techniques in particular have been recently shown to achieve impressive recognition performance (Goëau et al., 2017). Some of these results were integrated into effective web or mobile tools and have initiated close interactions between computer scientists and end-users such as ecologists, botanists, educators, land managers, and the general public. One remarkable realization in this domain is the Pl@ntNet mobile application (Affouard et al., 2017). It is used in an eponymous citizen science initiative (SciStarter, available at https://scistarter.com/project/16909-Plnt-Net) by a growing number of users around the world (more than 6 million downloads since 2013), and tens of thousands of plant pictures are submitted each day. Because a large fraction of this observation stream is geolocalized, it has great potential in terms of biodiversity monitoring and species distribution modeling (SDM). As the use of opportunistic data coming from citizen science initiatives has already been proven by Giraud et al. (2016) to strengthen the estimate of relative bird species abundance, we can expect other potential uses for such data types in a botanical context with Pl@ntNet.

Acquiring a large amount of opportunistic data still occurs at the expense of data quality and reliability, however. Many irrelevant pictures are submitted by the users of the Pl@ntNet application. This includes non-plant pictures, plant pictures of poor quality, or pictures of taxa that are not in the designated checklist (e.g., potted plants, ornamental and horticultural varieties, hybrids). Because the machine learning algorithm is not able to filter all of these pictures, many of them result in false positives (i.e., they are predicted as occurrences of species belonging to the checklist). Indeed, for a species automatically identified from a picture, two problems may induce identification error: (1) there is an intrinsic taxonomic uncertainty given the picture alone (i.e., it does not contain the discriminant visual pattern[s] that would make an expert certain about the exact species identification) or (2) the species was misidentified. Figure 1 illustrates typical examples of identification errors for *Acer monspessulanum* L. In Fig. 1B, one can see that the small symmetrical lobes at the base of the leaf might be confused with those of

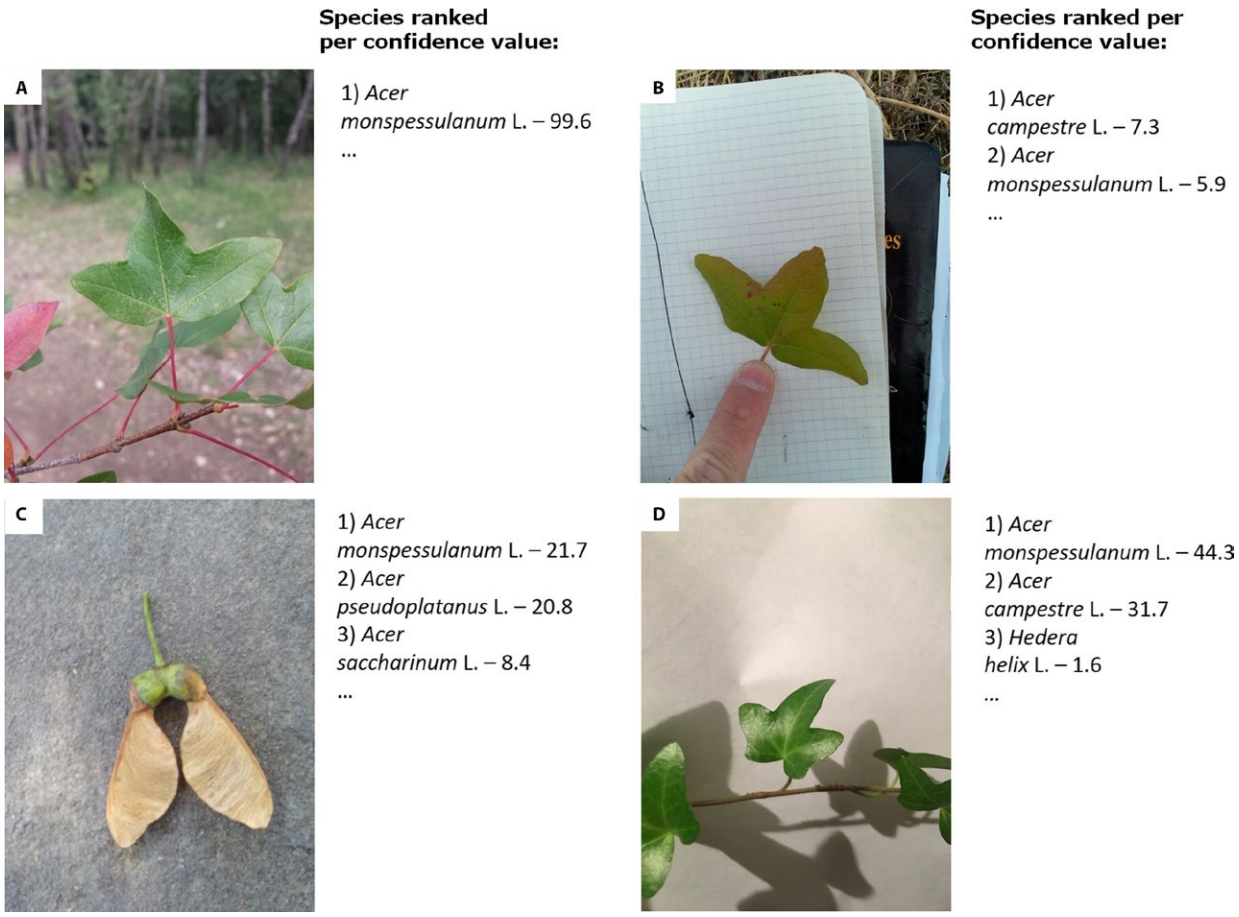

FIGURE 1. Four unvalidated Pl@ntNet plant pictures representing, or identified as, *Acer monspessulanum* and their respective predicted confidence values for the highest ranked species (the sum of scores over all species is always 100). (A) The species is *A. monspessulanum* and is well predicted. (B) The species is *A. monspessulanum*, but the model confounds it with *A. campestre*. (C) The species is *A. monspessulanum* or *A. pseudoplatanus*, but the species cannot be determined with the fruit only; there is an intrinsic taxonomic uncertainty. (D) The species is *Hedera helix* but is predicted as *A. monspessulanum* because this leaf is quite similar, as one can compare with (A).

a young specimen of *A. campestre* L., which is probably the cause of the model uncertainty. Figure 1C well illustrates the problem of taxonomic uncertainty, as several species cannot be distinguished by the feature recorded in the observer's image where there is high proximity of the confidence values of the first two species. Finally, Fig. 1D shows a leaf of *Hedera helix* L. with three major lobes that have strong visual similarity to those of the *A. monspessulanum* leaf. Manually cleaning such large and noisy data streams is not possible. These problems imply that all species are not equal in their potential for automatic identification. There are several factors that make a species automatically identifiable from a photograph: the scale of the discriminant visual pattern (for example, there are many issues with the Poaceae family because discriminant features are often too small to be easily captured with a photograph), the visual saliency of the pattern compared to other species, and the temporality of the pattern due to the phenology of its organ.

In this article, we explore the possibility of exploiting automatically identified observations, without human validation, for SDM. Specifically, we study the impact of the degree of uncertainty of the retained occurrences when training the popular MAXENT niche modeling approach (Merow et al., 2013). Given the type of Pl@ntNet users, candidate species have to be automatically identifiable by non-expert observers who are often not familiar with the discriminant part of the plant that needs to be photographed. In addition, species that are visually similar in pictures must be avoided, and the chosen species must be well illustrated in the predictive model training database. In addition to these criteria that allow automatic species identification, we must take into account the requirements using SDM on presence-only data to acquire meaningful results. More precisely, the species must have contrasted environmental preferences regarding the study domain, its realized habitat must not be overly constrained by its dispersal capacity or important historical perturbations, and there must be enough observation points regarding the environmental variables considered.

Considering these constraints on species selection, the available data, and the potential use-cases, we applied our protocol to the modeling of the distribution of five species classified in major and moderate categories of invasion by the National Mediterranean Botanical Conservatory of Porquerolles for the southeastern region of France (Conservatoire botanique national méditerranéen de Porquerolles, 2018). Invasive species represent a major economic cost to our society (estimated at nearly €12 billion a year in Europe) and are one of the main threats to biodiversity conservation (Weber and Gut, 2004). The early detection of the appearance of these species is a key element in managing them and reducing the cost of such management. The analysis of Pl@ntNet data can provide a highly valuable response to this problem because the presence of these species is often correlated with that of human activity (and thus to the density of Pl@ntNet data occurrences), and the constant flow of observations enables annual monitoring of species distributions.

METHODS

Automatic species identification and the Pl@ntNet workflow

We first present the workflow of the Pl@ntNet system that yields automatically identified observations. To compute automatic species identification, we use a convolutional neural network (CNN). CNNs have been shown to considerably improve the accuracy of automated plant species identification compared to previous methods (Grinblat et al., 2016; Ghazi et al., 2017; Goëau et al., 2017). More generally, CNNs recently received much attention in the computer vision community because of the impressive performance they can achieve on a large variety of classification tasks. Details of the CNN architecture and of the training procedure we used in this study are provided in Appendix 1. The network was trained in a supervised manner on a set of 332,000 humanly validated plant images belonging to approximately 11,000 species and an additional rejection class (containing non-plant pictures taken by Pl@ntNet users, e.g., faces, animals, manufactured objects). These species cover a large part of the European and North African floras, according to the network of people initially involved in the production and validation of these data (this network was initiated with the Tela Botanica non-governmental organization [http://www.tela-botanica.org] and the network of French-speaking botanists, composed of professionals and amateurs). This data set also includes a few hundred species of common tropical plants from two tropical regions: the Indian Ocean region and tropical Amazonia. Data from these two regions were collected by scientists and engineers from research institutes and universities working on these flora, representatives of the Tela Botanica network in these regions, and Pl@ntNet users. The data validation process was conducted using the IdentiPlante web tool (http://www.tela-botanica.org/appli:identiplante), essentially dedicated to the Tela Botanica community, and was also accessible on the Pl@ntNet Android app. These applications display all botanical records shared by the project members. Logged-in users are able to provide new identifications, post comments, and vote on previous identifications. The revised data are regularly crawled by the visual search engine, which picks up observations considered correctly identified according to a predefined set of rules on the votes and on possible conflicts. These validation tools allow coverage of a growing number of species, from 800 in 2013 up to 11,000 in 2016.

Species distribution modeling using automatically identified Pl@ntNet observations

We performed SDM based on the unvalidated Pl@ntNet observations made in France in 2016. In total, the data represent approximately 2 million observations (most observations have only one image and some have up to five images). Each image $\mathbf{x}$ was passed to the CNN to receive an automated species prediction in the form of a categorical distribution $p(k|\mathbf{x})$ estimating the probability that the image $\mathbf{x}$ is from the k-th species (according to the softmax classification layer of the CNN). For the observations composed of several images, the predictions were simply averaged (i.e., $p(k|x) = 1/n_x \cdot \sum p(k|\mathbf{x}_i)$ for an observation $\mathbf{x}$ composed of n_x images $\mathbf{x}_i$). We then kept only the observations for which the most probable species (denoted as k_{max}) belonged to the set of the five potential invasive species considered in our study: *Acer negundo* L., *Carpobrotus edulis* (L.) N. E. Br., *Erigeron karvinskianus* DC., *Opuntia ficus-indica* (L.) Mill., and *Reynoutria japonica* Houtt. The resulting number of occurrences per species and per interval of confidence values $p(k_{max}|x)$ is provided in Fig. 2. For low values of $p(k_{max}|x)$, the level of noise is important (e.g., with several false positives for $p(k_{max}|x) < 30\%$). For the highest values of $p(k_{max}|x)$ (e.g., $p(k_{max}|x) > 95\%$), the level of noise is more reasonable but the number of occurrences is also much lower. Thus, to maximize SDM performance, one could expect a positive trade-off with an intermediate threshold.

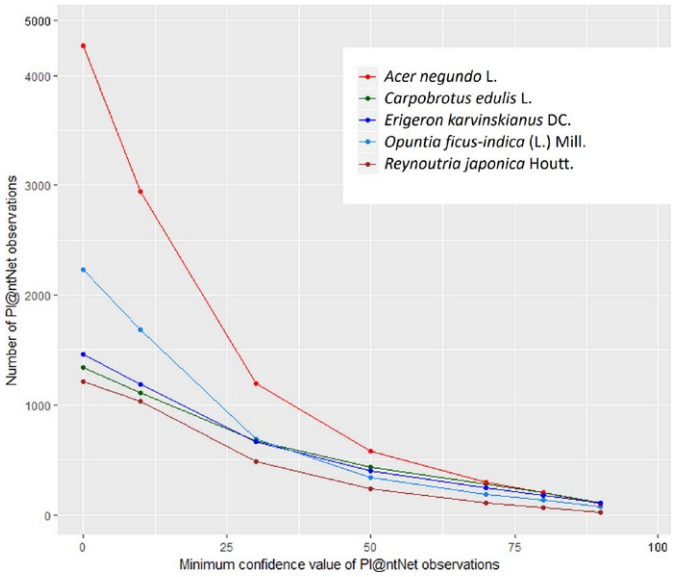

FIGURE 2. The number of Pl@ntNet observations per species and per confidence values $p(k_{max}|x)$.

To validate the species distribution models trained from automatically identified data, we used a second reference data set comprising count data collected and validated by French expert naturalists. This data set, referred to as Inventaire National du Patrimoine Naturel (INPN; https://www.gbif.org/dataset/75956ee6-1a2b-4fa3-b3e8-ccda64ce6c2d; Dutrève and Robert, 2016), comes from the Global Biodiversity Information Facility (https://www.gbif.org/). The underlying occurrences were collected in various contexts, including floras and regional catalogs, specific inventories, field notebooks, and surveys carried out by botanical conservatories. We kept only a subset of these data corresponding to the five invasive species considered in our study. The resulting data set contains 20,810 occurrences (see Table 1 for the detailed numbers per species) aggregated in 3242 quadrat cells of 100 km^2 distributed on a regular grid of 5175 quadrat cells covering the French territory.

Species distribution models were computed via MAXENT (Phillips et al., 2004, 2006), a popular environmental niche modeling method. In particular, we used the implementation of the *maxnet* (Phillips et al., 2017) R package that expands the input environmental variables with several functions (including linear, quadratic, threshold, hinge, and first-order interactions). Because we used presence-only SDM, we used pseudo-absence localities for model parameterization (see Appendix 2 for more details). MAXENT was computed on a set of 29 input environmental variables, including bioclimatic, pedological, topological, hydrographical, and land cover variables from CHELSA Climate data 1.1 (Karger et al., 2017), Consultative Group on International Agricultural Research–Consortium for Spatial Information (CGIAR-CSI) potential evapo-transpiration (ETP) data (Zomer et al., 2007, 2008), ESDBv.2 (Panagos, 2006; Van Liedekerke et al., 2006; Panagos et al., 2012), U.S. Geological Survey Digital Elevation data, the Institut National de l'information Géographique et

TABLE 1. Detailed number of occurrences in the Inventaire National du Patrimoine Naturel (INPN) data set by species.

Species name	No. of observations	No. of 100-km^2 areas
Acer negundo L.	5217	904
Carpobrotus edulis (L.) N. E. Br.	484	114
Erigeron karvinskianus DC.	711	306
Opuntia ficus-indica (L.) Mill.	120	44
Reynoutria japonica Houtt.	14,278	2623

TABLE 2. List and details of the environmental descriptors used in this study.

Name	Description	Nature	Values[a]	Local image
CHBIO_2	Mean monthly temp (max, min)	quanti.	[7.8, 21.0]	Yes
CHBIO_7	Temp. annual range	quanti.	[16.7, 42.0]	Yes
CHBIO_8	Mean temp. of wettest quarter	quanti.	[−14.2, 23.0]	Yes
CHBIO_9	Mean temp. of driest quarter	quanti.	[−17.7, 26.5]	Yes
CHBIO_10	Mean temp. of warmest quarter	quanti.	[−2.8, 26.5]	Yes
CHBIO_11	Mean temp. of coldest quarter	quanti.	[−17.7, 11.8]	Yes
CHBIO_13	Precip. of wettest month	quanti.	[43.0, 285.5]	Yes
CHBIO_14	Precip. of driest month	quanti.	[3.0, 135.6]	Yes
CHBIO_15	Precip. seasonality (CV)	quanti.	[8.2, 26.5]	Yes
CHBIO_18	Precip. of warmest quarter	quanti.	[19.8, 851.7]	Yes
CHBIO_19	Precip. of coldest quarter	quanti.	[60.5, 520.4]	Yes
etp	Potential evapotranspiration	quanti.	[133, 1176]	Yes
alti	Elevation	quanti.	[−188, 4672]	Yes
shade	Shade level	quanti.	[0, 1]	No
slope	Ground slope	quanti.	[0, 13457]	No
dmer	Distance to coastline	quanti.	[|0, 32767|]	No
droute	Distance to roads	quanti.	[|0, 32767|]	No
proxi_eau	<50 m to fresh water	bool.	{0, 1}	Yes
awc_top	Topsoil available water capacity	ordinal	{0, 120, 165, 210}	Yes
bs_top	Base saturation of the topsoil	ordinal	{35, 62, 85}	Yes
cec_top	Topsoil cation exchange capacity	ordinal	{7, 22, 50}	Yes
crusting	Soil crusting class	ordinal	[|0, 5|]	Yes
dgh	Depth to a gleyed horizon	ordinal	{20, 60, 140}	Yes
dimp	Depth to an impermeable layer	ordinal	{60, 100}	Yes
erodi	Soil erodibility class	ordinal	[|0, 5|]	Yes
oc_top	Topsoil organic carbon content	ordinal	{1, 2, 4, 8}	Yes
pd_top	Topsoil packing density	ordinal	{1, 2}	Yes
text	Dominant surface textural class	ordinal	[|0, 5|]	Yes
clc	Ground occupation	categ.	[|1, 48|]	Yes

Note: bool. = Boolean data; categ. = categorical data; CV = coefficient of variation of monthly precipitation; quanti. = quantitative data.

[a]Data presented in curly brackets ({ }) contain the list of all possibles values of the variable, i.e., a discrete ensemble; square brackets ([ ]) indicate the continuous range of values that can take the variable, i.e., a continuous interval; vertical lines indicate the range of integers between the two bounds given, i.e., a discrete interval.

forestière–Système d'Administration Nationale des Données et Référentiels sur l'Eau (IGN-SANDRE) BD Carthage, CORINE Land Cover 2012 data, and IGN ROUTE500 data. The detailed methodology of how these variables were collected and formatted is described in Appendix 3. The full list of the variables used is presented in Table 2. For each of the considered species, we computed seven models with varying levels of minimal confidence of species occurrences, i.e., different threshold values $p_{min}(k_{max}|x)$ of the categorical probability $p(k_{max}|x)$. We know that the global sampling effort in Pl@ntNet is highly correlated with human population density and the proximity to roads and to the coastline. In our study, the sampling intensity was so high compared to the species abundance that we strongly overestimated the species abundance in cities, on beaches, and on roads. Consequently, we fitted MAXENT models, including variables of urban areas, proximity to roads, and distance to the coastline. In the predicted abundance function, we then kept these variables constant across space to cancel the effect of the sampling effort (see Appendix 2 for more details). This approach has already been proposed and successfully used in the literature of SDMs (Warton et al., 2013; Stolar and Nielsen, 2015). The predictive effectiveness of the models was then assessed using the INPN count data as a validation set. We used two evaluation metrics: (1) the true skills statistics (TSS) equal to the sum of the sensitivity and the specificity minus one (as described in Allouche et al., 2006), and (2) the accuracy on 10% densest quadrats (A10DQ; see Appendix 2 for more details). The TSS is the sum of sensibility and specificity minus one when comparing the SDM predicted presences/absences of a species with the references (the INPN data set). It is a meaningful measure to evaluate the model's ability to detect presences while simultaneously minimizing false positives. It is computed through binarization of SDM continuous prediction based on the threshold that maximizes the TSS. We chose the A10DQ as a complementary metric because it evaluates the accuracy of the models in predicting the quadrats with the highest abundance (INPN count), which is an especially interesting property from the perspective of invasive species management.

RESULTS

Figure 3 displays the evaluation metrics as a function of the confidence threshold $p_{min}(k_{max}|x)$ applied to filter the automatic predictions. We found that the confidence threshold had variable influence depending on the species, but there was an overall trend represented by the average curve (Fig. 3, black solid line). Too-low thresholds did not allow for filtering identification errors sufficiently, thus the model was biased by the presence of too many irrelevant occurrences. A too-high threshold (above 70%) also degraded the model performance (in particular, the accuracy of the quadrat cells with the higher level of counts; see Fig. 3) because the number of retained occurrences in the training set decreased significantly with increasing threshold. Models based on too few occurrences could not provide a relevant prediction of species distribution. With the current Pl@ntNet data, the chosen species, and the variables, a confidence threshold of 70% represented a good compromise for SDM. It filtered identification errors effectively for most species while retaining enough occurrences for model training. The most problematic species was *Reynoutria japonica*: it had very poor TSS for all thresholds (a TSS score of 0 would be a random prediction of presence and absence), indicating that the SDM did not distinguish presence and absence zones very well. This species is the most widespread, which leads to poor SDM performances. Nevertheless, for the best threshold, A10DQ showed that 20% of the densest INPN quadrats were predicted by the model fitted on Pl@ntNet, which is significantly better than a random ranking of quadrats (which would give an average of 10% and a standard deviation of 1.3%). Consequently, the model could capture information on the distribution of *Reynoutria* from the Pl@ntNet data. Conversely, very good results were obtained for both metrics for *Opuntia ficus-indica* and *Carpobrotus edulis*.

Figure 4 further shows the distributions predicted for each species using $p_{min}(k_{max}|x) = 70\%$. For comparison, we also displayed the expert count data of INPN, as well as the specificity and sensitivity of our model measured with that data (at TSS max). Most regions with high INPN counts were reasonably well predicted by

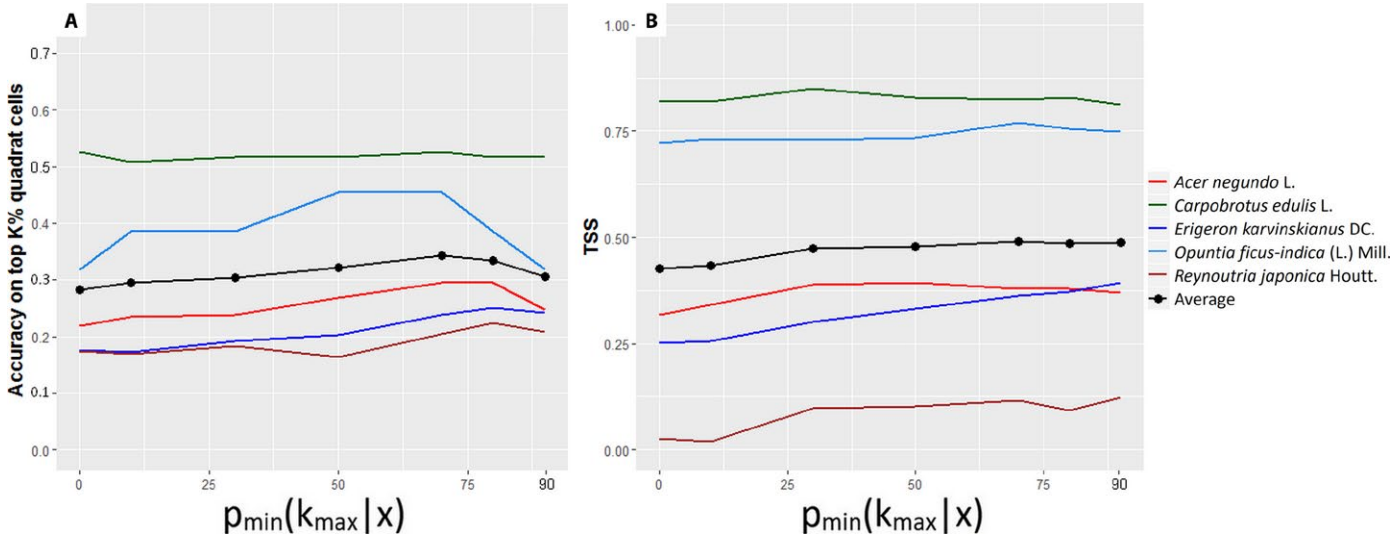

FIGURE 3. Predictive effectiveness of the species distribution models trained on Pl@ntNet data as a function of the confidence threshold value $p_{min}(k_{max}|x)$ showing accuracy on the 10% densest quadrats (A) and true skill statistics (TSS; conversion of prediction value into presence/absence with the threshold that maximizes TSS) (B).

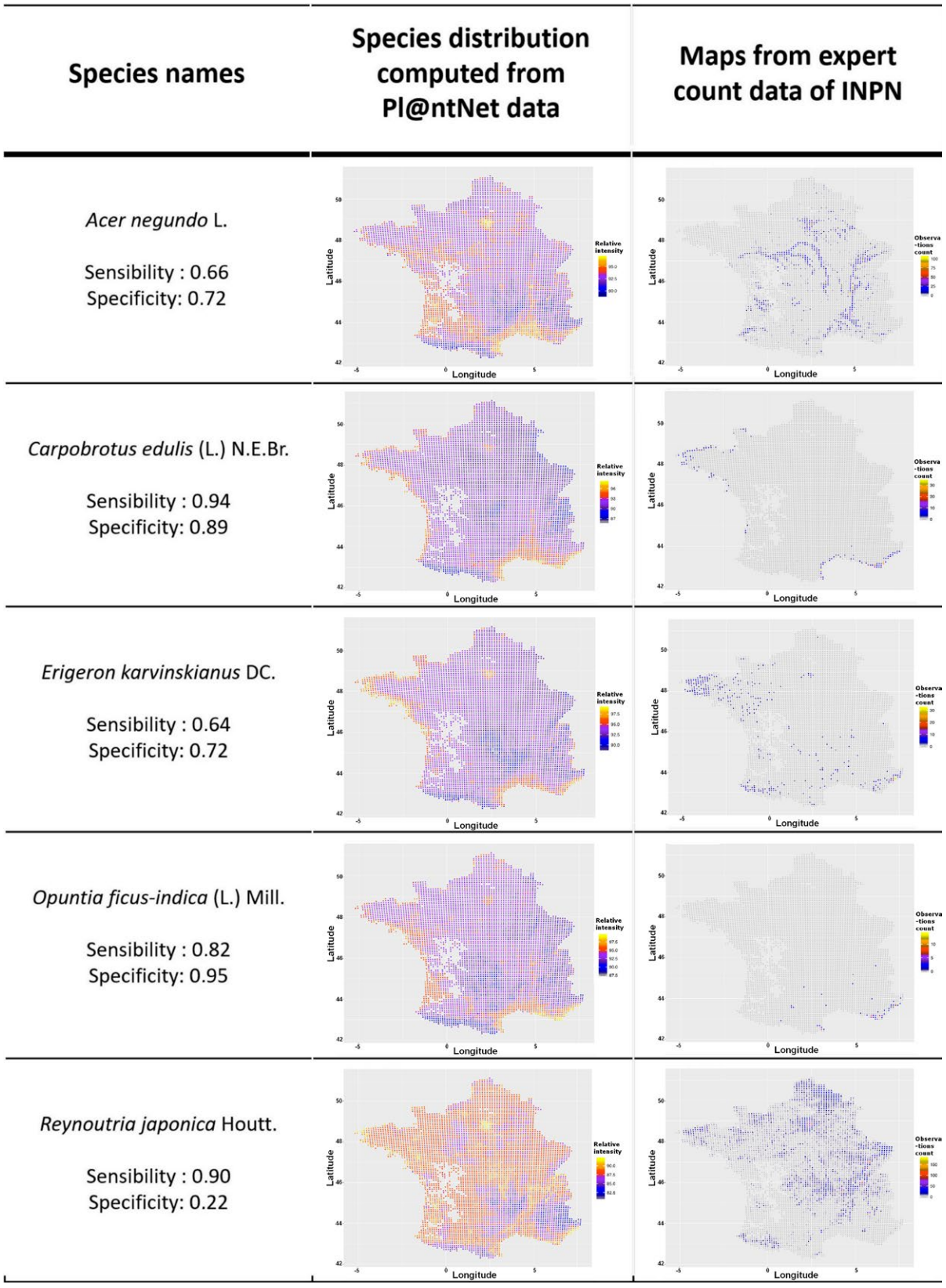

FIGURE 4. Maps of species distribution models computed from Pl@ntNet data (based on $p_{min}(k_{max}|x) = 70\%$) and of expert count data from the Inventaire National du Patrimoine Naturel (INPN). The sensibility and specificity used for the computation of the true skill statistics (for pmin(kmax|x) = 70) is provided for each species.

the models. Accordingly, sensitivity values were generally accurate for most species. Nevertheless, there were also regions for which the Pl@ntNet model and INPN data disagreed; in these regions the Pl@ntNet model predicted high abundances but there were none or very few occurrences in the INPN data. The strongest disagreement occurred for *Reynoutria japonica*, i.e., the taxon for which the specificity was the lowest. Other false-positive prediction regions included the west coast for *Opuntia ficus-indica* and *Carpobrotus edulis* and the "Golfe du Lion" (arc on the southeast coast) for *O. ficus-indica* and *Erigeron karvinskianus*.

DISCUSSION

Visual inspection of Pl@ntNet observations occurring in such false-positive regions revealed that for the vast majority such observations did not correspond to erroneous identifications ($p_{min}(k_{max}|x)$ = 70% is a high enough threshold to remove noise efficiently). Rather, they corresponded to real occurrences that can be classified in three main categories (see Fig. 5 for examples of observations belonging to the different categories). The first category can be qualified as cultivated specimens, i.e., specimens planted and/or maintained by humans such as gardening plants, house plants, ornamental plants in city parks, etc. Most occurrences of *Opuntia ficus-indica* on the west coast belonged to this category. A second category of observations could be qualified as casual invasive specimens, i.e., isolated specimens that often flourish close to human construction but that do not form self-replacing populations. Cultivated and casual invasive specimens present in the observations reveal that the species is able to grow in a great diversity of habitats. These specimens should be identified, either to (1) filter them for model learning, (2) evaluate the correlation between species gardening intensity and its abundance in wild surroundings, or (3) learn more complex models that integrate dispersal mechanisms and quantify more precisely the importance of gardening intensity on the species' capacity to colonize a region. To identify cultivated specimens, several options are possible: for example, learning models can be used to identify the context of the picture or the user can be asked to clarify the type of environment where the observation was made, especially when observations appear ambiguous. Apart from the issue of correctly predicting species occurrences in the wild, frequent occurrences of cultivated and casual invasive specimens in a region where there is no presence in the wild can reflect the risk of future invasion in the wild.

A last category of observations can be qualified as newly inventoried invasive specimens, i.e., non-isolated specimens living in natural areas that have yet to be inventoried in the INPN data. Notably, the majority of occurrences of *Carpobrotus edulis* on the west coast belong to this category. Newly inventoried invasive specimens could provide an early warning for territory managers. For

FIGURE 5. Pl@ntNet observations with a species prediction score of more than 70% for plants living in natural conditions or cultivated for ornamental purpose.

example, we found newly inventoried specimens of *Reynoutria japonica* in the Pl@ntNet data, and we suspect that poor performance of its SDM could reflect a negative bias in the evaluation metrics of this species. Typically, specimens occurring outside of presence areas identified by experts and not categorized as cultivated or casual invasive should be prioritized for expert validation.

In this study, our sampling effort correction approach was based on prior knowledge of sampling intensity in the Pl@ntNet data. We could not evaluate the errors related to the sampling effort bias without complementary systematic survey data. Nevertheless, the INPN data have their own heterogeneity in the spatial distribution of the sampling effort. These data were collected by independent regional conservatories, and variations in sampling by different workforces may have introduced regional heterogeneity. Furthermore, some zones are not surveyed by conservatories, typically cities in most cases, which tends to bias the Pl@ntNet model error in urban areas. The study of global sampling effort bias is crucial for exploiting presence-only data collected without protocol. The spatially heterogeneous sampling effort is especially problematic when it is correlated with environmental variables impacting the species distribution. For example, the sampling effort is correlated with the distance to the coastline, which is also a variable influencing the abundance of *Opuntia ficus-indica*, *Erigeron karvinskianus*, and *Carpobrotus edulis*. Because our bias correction method removes the distance to the coastline effect, it partially removes the ability of the model to capture this effect on the species distribution. When we included these variables in the predicted distribution of the three species (results not presented in this article), we found a much greater predicted abundance gradient toward the coast. However, the maps presented in Fig. 4 show that the model captured a part of the coastal effect through other variables that are correlated with the distance to coastline. The same problem will occur with other invasive species that tend to grow near roads as a result of constant perturbation or dispersal mechanisms. More generally, we note that the presence of invasive species is strongly influenced by human activity. It is also highly correlated with observational intensity in opportunistic presence-only data. Thus, this category of species represents a major methodological challenge for improving SDM based on presence-only data and represents a clear path for future research.

CONCLUSIONS

This study is the first to evaluate the potential of automated identification of opportunistic plant observations for modeling species distributions. The described methodology allowed us to analyze the potential usefulness of the Pl@ntNet data. By comparing SDMs trained on Pl@ntNet unvalidated observations with validated independent count data on a large spatial scale, we found that the data are rich enough to be used for SDM with only a single year of data collection. However, we also showed that distributions reported from Pl@ntNet data do not precisely match those of expert data. The main reasons for these deviations appear to be the presence of cultivated or casual invasive specimens in the data set, the detection of real presence in new areas, and the limits of the sampling bias correction method. Noticing these limits allowed us to underline significant research challenges for SDMs and to provide possible methods to usefully integrate information provided by opportunistic citizen science observations into conservation management.

ACKNOWLEDGMENTS

The authors thank the Inventaire National du Patrimoine Naturel (INPN) and the Fédération française des Conservatoires botaniques nationaux for access to the expert count data used in this study. We also would like to thank the Tela Botanica and Pl@ntNet community who have contributed to produce and revise the data used in this study.

LITERATURE CITED

Affouard, A., H. Goëau, P. Bonnet, J. C. Lombardo, and A. Joly. 2017. Pl@ntnet app in the era of deep learning. *In* 5th International Conference on Learning Representations, 24–26 April 2017, Toulon, France.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223–1232.

Carranza-Rojas, J., H. Goëau, P. Bonnet, E. Mata-Montero, and A. Joly. 2017. Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology* 17: 181.

Casanova, D., J. J. de Mesquita Sá Junior, and O. M. Bruno. 2009. Plant leaf identification using Gabor wavelets. *International Journal of Imaging Systems and Technology* 19: 236–243.

Champ, J., T. Lorieul, P. Bonnet, N. Maghnaoui, C. Sereno, T. Dessup, J. M. Boursiquot, et al. 2016. Categorizing plant images at the variety level: Did you say fine-grained? *Pattern Recognition Letters* 81: 71–79.

Conservatoire botanique national méditerranéen de Porquerolles. 2018. Espèce végétale exotique envahissante (EVEE) [online]. Website http://www.invmed.fr/src/listes/index.php?idma=33 [accessed 31 January 2018].

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *In* Proceedings, Institute of Electrical and Electronics Engineers (IEEE) Computer Society Conference on Computer Vision and Pattern Recognition, 20–25 June 2009, 248–255. IEEE, Piscataway, New Jersey, USA.

Dutrève, B., and S. Robert. 2016. INPN (Inventaire National du Patrimoine Naturel): Données flore des Conservatoires botaniques nationaux (CBN) agrégées par la Fédération des Conservatoires botaniques nationaux (FCBN). Version 1.1. Service du Patrimoine naturel (SPN), Muséum national d'Histoire naturelle, Paris, France. Occurrence Dataset https://doi.org/10.15468/omae84 via GBIF.org [accessed 30 August 2017].

Fithian, W., and T. Hastie. 2013. Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics* 7: 1917.

Gaston, K. J., and M. A. O'Neill. 2004. Automated species identification: Why not? *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 359: 655–667.

Ghazi, M. M., B. Yanikoglu, and E. Aptoula. 2017. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* 235: 228–235.

Giraud, C., C. Calenge, C. Coron, and R. Julliard. 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics* 72: 649–658.

Goëau, H., P. Bonnet, and A. Joly. 2016. Plant identification in an open-world (LifeCLEF 2016). *In* K. Balog, L. Cappellato, N. Ferro, and C. Macdonald [eds.], Working notes of CLEF 2016–Conference and labs of the evaluation forum, 5–8 September 2016, Évora, Portugal. *CEUR Workshop Proceedings* 1609: 428–439.

Goëau, H., P. Bonnet, and A. Joly. 2017. Plant identification based on noisy web data: The amazing performance of deep learning (LifeCLEF 2017). *In* L. Cappellato, N. Ferro, L. Goeuriot, and T. Mandl [eds.], Working notes of CLEF 2017–Conference and labs of the evaluation forum, 11–14 September 2017, Dublin, Ireland. *CEUR Workshop Proceedings* 1866: ceur-ws.org/Vol-1866/invited_paper_9.pdf.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep learning, vol 1. MIT Press, Cambridge, Massachusetts, USA.

Grinblat, G. L., L. C. Uzal, M. G. Larese, and P. M. Granitto. 2016. Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture* 127: 418–424.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In* Proceedings, IEEE Computer Society Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015, 1026–1034. IEEE, Piscataway, New Jersey, USA.

Ioffe, S., and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In* Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015. *PMLR* 37: 448–456.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *In* Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, Florida, USA, 3–7 November 2014, 675–678. ACM, New York, New York, USA.

Joly, A., P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, et al. 2016. A look inside the Pl@ntNet experience. *Multimedia Systems* 22: 751–766.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, et al. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data* 4: 170122.

Lee, S. H., C. S. Chan, P. Wilkin, and P. Remagnino. 2015. Deep-plant: Plant identification with convolutional neural networks. *In* Proceedings, Institute of Electrical and Electronics Engineers (IEEE) International Conference on Image Processing (ICIP), Macau, China, 16–18 September 2015, 452–456. IEEE, Piscataway, New Jersey, USA.

Merow, C., M. J. Smith, and J. A. Silander. 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* 36: 1058–1069.

Panagos, P. 2006. The European soil database. *GEO: Connexion* 5: 32–33.

Panagos, P., M. Van Liedekerke, A. Jones, and L. Montanarella. 2012. European Soil Data Centre: Response to European policy support and public data requirements. *Land Use Policy* 29: 329–338.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. *In* Proceedings of the Twenty-First International Conference on Machine Learning, New York, New York, USA, 10–14 June 2004, p. 83. ACM Digital Library, New York, New York, USA.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.

Phillips, S. J., R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. 2017. Opening the black box: An open-source release of Maxent. *Ecography* 40: 887–893.

Stolar, J., and S. E. Nielsen. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions* 21: 595–608.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, et al. 2015. Going deeper with convolutions. *In* Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 7–12 June 2015, 1–9. IEEE, Piscataway, New Jersey, USA.

Van Liedekerke, M., A. Jones, and P. Panagos. 2006. ESDBv2 Raster Library: A set of rasters derived from the European Soil Database distribution v2. 0. European Commission and the European Soil Bureau Network, CDROM, EUR, 19945.

Wäldchen, J., and P. Mäder. 2017. Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-016-9206-z.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS One* 8: e79168.

Weber, E., and D. Gut. 2004. Assessing the risk of potentially invasive plant species in central Europe. *Journal for Nature Conservation* 12: 171–179.

Wilf, P., S. Zhang, S. Chikkerur, S. A. Little, S. L. Wing, and T. Serre. 2016. Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences USA* 113: 3305–3310.

Yanikoglu, B., E. Aptoula, and C. Tirkaz. 2014. Automatic plant identification from photographs. *Machine Vision and Applications* 25: 1369–1383.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? *In* Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger [eds.], Proceedings of Neural Information Processing Systems (NIPS 2014), Montréal, Canada, 8–13 December 2014. *Advances in Neural Information Processing Systems* 27: 3320–3328.

Zomer, R. J., D. A. Bossio, A. Trabucco, L. Yuanjie, D. C. Gupta, and V. P. Singh. 2007. Trees and water: Smallholder agroforestry on irrigated lands in Northern India. IWMI Research Report 122. International Water Management Institute (IWMI), Colombo, Sri Lanka.

Zomer, R. J., A. Trabucco, D. A. Bossio, and L. V. Verchot. 2008. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems and Environment* 126: 67–80.

APPENDIX 1. Detailed architecture and training procedure of the convolutional neural network used to compute the automated identifications.

The main strength of convolutional neural network (CNN) technologies comes from their ability to learn discriminant visual features directly from the raw pixels of the images without exponentially increasing the model variables as the dimensionality grows (Goodfellow et al., 2016). This is achieved by stacking multiple convolutional layers, i.e., the core building blocks of a CNN. In general, a convolutional layer takes images as input and produces as output feature maps corresponding to different convolution kernels while looking for different visual patterns.

To get to specific choices in the architecture, we used an extended version of the GoogleNet model (Szegedy et al., 2015) that is a very deep CNN that stacks several so-called inception layers. As in Carranza-Rojas et al. (2017), we extended the base version with batch normalization (Ioffe and Szegedy, 2015), which has been proven to speed up convergence and limit overfitting, and with a parametric rectified linear unit (PReLU) activation function (He et al., 2015) instead of the traditional rectified linear unit (ReLU).

To improve the generalization ability of the network, we used transfer learning, which is a powerful paradigm to overcome the lack of sufficient domain-specific training data. Deep learning models have to be trained on thousands of pictures per class to converge on accurate classification models. It has been shown that the first layers of deep neural networks deal with generic features (Yosinski et al., 2014) so that they are generally usable for other computer vision tasks. Consequently, they can be trained on arbitrary training image data. The last layers contain more or less generic information transferable from one classification task to another. These layers are expected to be more informative for the optimization algorithm than a random initialization of the weights of the network. Therefore, a common practice is to initialize the network by pre-training it on a large available data set and then fine-tune it on the scarcer domain-specific data. Many networks are pre-trained on the generalist data set ImageNet (Deng et al., 2009), which covers a large variety of visual concepts, including animals, vehicles, and manufactured objects. Because the GoogleNet model we used was already pre-trained on this generalist data set, we used the following methodology for fine-tuning it on our data set of 11,000 species (using the Caffe framework [Jia et al., 2014]):

1. The linear classification layer was replaced by a new one aimed at classifying the new classes (i.e., the 11,000 species). It was initialized with random weights and the learning rate was multiplied by 10 for this layer.
2. The other layers were kept unchanged to initialize the network with the weights learned from ImageNet.
3. The network was trained on the 332,000 plant images of our training set.

A batch size of 16 images was used for each iteration, with a learning rate of 0.0075 with images of 224 × 224 resolution. Simple crop and resize data augmentation was used with the default settings of the Caffe framework.

APPENDIX 2. Description of Pl@ntNet data post-treatments, generation of quadrature points, and experimental procedure. Results were obtained using R.

Filtering of Pl@ntNet geolocated observations: We used the unvalidated observations collected by Pl@ntNet users during the year 2016. We kept only observations for which one of our five species was ranked first according to the identification score. We first selected those whose GPS geolocation falls in the French Metropolitan territory (polygon: getData(country="FRA",level=0), function from package *raster*) excluding Corsica, or are closer than 500 m to the coastline (because of coordinate error). Because observations are very often duplicated due to a repeated submission of the same set of pictures, we kept only one of the identical observations. Unsatisfactory automatic identification of the same specimen allowed the user to take new pictures of the specimen and submit it again. This kind of duplication was removed by the following procedure: for two occurrences closer than 60 sec in time and 100 m in space, we kept the one with highest $p(k_{max}|x)$.

Quadrature points: MAXENT can be interpreted as a non-homogeneous Poisson process model (Fithian and Hastie, 2013). Thus, computing a MAXENT model from observations requires integration of its intensity function over the spatial domain of study D (in this study, the French territory). For this purpose, it approximates the integral with quadrature points, also called "pseudo-absences," that represent the distribution of the environmental descriptors on D. As our domain was wide, and some of our descriptors vary with high spatial frequency (like distance to roads or proximity to fresh water), we used a high number of quadrature points. We generated 101,632 points on a grid with a similar spacing of 0.025 in longitude (approximately 2 km) and latitude (approximtely 2.8 km), and strictly included in the French polygon (see above).

Prediction of model relative abundance for a plot and attribution of quadrature points to plots: With a fitted MAXENT model, we can evaluate its intensity function at every quadrature point via environmental descriptors, which gives a high-resolution map of predicted relative abundance across France. This fine-resolution prediction includes the effect of high-frequency variables. However, to compare model predictions to counts on quadrat cells, we need to upscale our prediction: according to the properties of the inhomogeneous Poisson process, the law of the number of points falling in a quadrat cell is a Poisson law whose parameter is the integral of the intensity function over the quadrat cell. Because the quadrature points are regularly spaced, we can approximate this integral up to a factor (common to every quadrat cell because they have the same area) with the mean of intensity values over quadrature points contained in the quadrat cell. For some cells located mainly above sea or ocean, some did not contain any quadrature points, thus we attributed the closest one while removing it from its original plot. In this way, quadrat cells contained an average of 17.1 quadrature points.

Bias-corrected model prediction: We know that there is sampling bias in the Pl@ntNet observation data. The most important is high sampling effort in cities, close to roads, and near coastlines (because of use during tourist activities). In addition, we know that for the species of interest, distance to roads and cities has no strong link to real abundance. Because we want to remove the artificial importance of those variables in the concentration of observations, one strategy is to integrate the sampling variables in the intensity function, as is now commonly done in such cases (Warton et al., 2013). If there is no perfect linear link between sampling and abundance variables, we will correctly infer our abundance model. Finally, we predict an unbiased relative abundance by setting the sampling variables to a constant value everywhere in space. However, we cannot do this for the distance to coastline because this variable plays a key role in the real abundance of *Carpobrotus edulis*, *Opuntia ficus-indica*, and *Erigeron karvinskianus*.

Evaluation metric: The evaluation metric represents the proportion of the top 10% quadrats in terms of real count that are also in the top 10% in terms of model prediction. However, we have to define the last quadrat cell ranked in the top 10% for counts, which is problematic for some species because of ex aequo cells. That is why we defined the following procedure that is adjusted for each species in the percentage of top cells such that the metrics can be calculated and the percentage is the closest to 10%. It is known as accuracy on the 10% densest quadrats (A10DQ):

$$\frac{N_{p\&c}(i)}{N_c(i)}$$

Where $N_{p\&c}(i)$ is the number of cells that are contained in the $N_c(i)$ higher cells both in terms of count and of model prediction.

Calculation of $N_c(i)$: We order the cells by decreasing the count of **i** and note C_k the count of the k-th cell in this order. As we are interested in the quadrat cells ranked in the highest 10%, if $C_{518} > C_{519}$, we set $N_c(i) = 518$. Otherwise, $C_{518} = C_{519}$ (ex aequo exists for 518th position), then we note *sup* the position of the last cell with count C_{519} and *inf* the position of the first cell with count C_{519}. The chosen rule is to take $N_c(i)$ such that $N_c(i) = \text{Min}(|sup\text{-}518|, |inf\text{-}518|)$.

APPENDIX 3. Detailed methodology of how environmental variables were collected and formatted in our study.

We used data covering the French metropolitan territory, freely available on the web. The environmental descriptors are listed in Table 2. Because the original coordinate systems of the layers used varied among sources, we systematically converted them to WGS84

using the *rgdal* package in R, which was the reference coordinate system for our observations, quadrature points, and quadrat cells. In the following points, we describe the sources, nature, and eventual transformations of those environmental data:

- CHELSA Climate data 1.1: These are raster data with worldwide coverage and 1-km resolution. A mechanistic climatic model is used to make spatial predictions of monthly mean-max-min temperatures, mean precipitations, and 19 bioclimatic variables that are downscaled with statistical models integrating historical measures of meteorologic stations from 1979 to the present (see Karger et al., 2017). The data are under Creative Commons Attribution 4.0 International License (available at http://chelsa-climate.org/downloads/).
- The ESDB v2, 1kmx1km Raster Library (Panagos, 2006; Van Liedekerke et al., 2006; Panagos et al., 2012): The library contains multiple soil pedological descriptor raster layers covering Eurasia at a resolution of 1 km. We selected 10 descriptors from the library. They represent quantitative physico-chemical quantities of the soil (from the PedoTransfer Rules Database [PTRDB attributes, available at https://esdac.jrc.ec.europa.eu/content/ptrdb-attributes]) that have been deduced from soil classification with expert rules, and their values are aggregated in intervals. As there are few possible intervals by variables (2–6), we integrated them as categorical variables in MAXENT. The data are maintained and distributed freely for scientific use by the European Soil Data Centre at http://eusoils.jrc.ec.europa.eu/content/european-soil-database-v2-raster-library-1kmx1km.
- CORINE Land Cover 2012, version 18.5.1, 12/2016: This is a raster layer describing soil occupation with 48 categories across Europe (25 countries) at a resolution of 100 m. This classification is the result of an interpretation process applied to the earth's surface with high-resolution satellite images. We set this variable as categorical in MAXENT with only 30 relevant categories for our purposes. This database of the European Union is freely accessible online at: http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012.
- CGIAR-CSI ETP data: The Consultative Group on International Agricultural Research–Consortium for Spatial Information (CGIAR-CSI) distributes this worldwide monthly potential evapo-transpiration raster data. It is pulled from a model developed by Antonio Trabucco (Zomer et al., 2007, 2008). Rasters are estimated by the Hargreaves formula using mean monthly surface temperatures and standard deviation from WorldClim 1:4 (http://www.worldclim.org/version1), and radiation on top of atmosphere. The raster is at a 1-km resolution and is freely downloadable for a nonprofit use at http://www.cgiar-csi.org/data/global-aridity-and-pet-database#description.
- U.S. Geological Survey Digital Elevation data: The Shuttle Radar Topography Mission achieved in 2010 by the Endeavour shuttle measured digital elevation at 3 arcs per second resolution over most of the earth's surface. Raw measures have been post-processed by the National Aeronautics and Space Administration and the National Geospatial-Intelligence Agency to correct detection anomalies. This gives a precision measurement of approximately 90 m for this variable. The data are available from the U.S. Geological Survey and are downloadable on the EarthExplorer (https://earthexplorer.usgs.gov/). See https://lta.cr.usgs.gov/SRTMVF for more information.
- BD Carthage v3: BD Carthage is a spatial database holding information on the structure and nature of the French Metropolitan hydrological network. We focus on the geometric segments representing watercourses, polygons representing hydrographic fresh surfaces, and the ocean. The data have been produced by the Institut National de l'information Géographique et forestière (IGN) from an interpretation of the BD Ortho IGN. The database is maintained by SANDRE under free license for non-profit use and is downloadable at: http://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FXX/2014/arcgis/.

For "proxi_eau," i.e., the proximity to fresh water, we used QGIS (https://qgis.org/) to rasterize to a 12.5-m resolution, with a buffer of 50 m, (1) the shapefile COURS_D_EAU.shp and (2) the polygons of SURFACES_HYDROGRAPHIQUES.shp with attribute NATURE="Eau douce permanente". We then created the maximum of the proximity raster derived from COURS_D_EAU.shp and SURFACES_HYDROGRAPHIQUES.shp (so the value of 1 corresponds to an approximate distance of less than 50 m to a watercourse or hydrographic surface of fresh water). For "dmer," i.e., the distance to the ocean, we calculated, using QGIS, the distance raster at a resolution of 12.5 m to polygons with attribute TYPE="Pleine mer" in the shapefile SURFACES_HYDROGRAPHIQUES.shp of BD Carthage up to a distance of 32,767 m for storage format convenience.

- ROUTE500 1.1: This database register classifies road linkages between cities (highways, national roads, and departmental roads) in France in shapefile format, representing approximately 500,000 km of roads. It is produced under free license (all uses) by the IGN. Data are available online at http://osm13.openstreetmap.fr/~cquest/route500/. For deriving the variable "droute," the distance to the main roads networks, we used a similar procedure as for "dmer," calculating the distance raster for all the elements of the shapefile ROUTES.shp (segments).